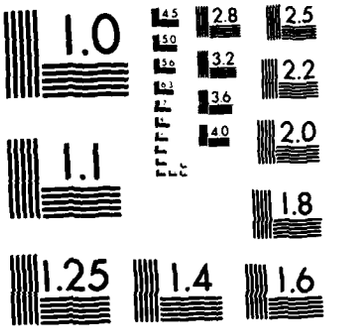
AD-A147 083 A COMPLETE CHARACTERIZATION OF TRIPLY BALANCED MATRICES 1/1
WITH APPLICATIONS.. (U) ILLINOIS UNIV AT CHICAGO CIRCLE
STATISTICAL LAB A HEDAYAT ET AL. AUG 84 TR-84-5
UNCLASSIFIED AFOSR-TR-84-0870 AFOSR-80-0170 F/G 12/1 NL



A COMPLETE CHARACTERIZATION OF TRIPLY BALANCED MATRICES WITH APPLICATIONS TO SURVEY SAMPLING

BY

A. HEDAYAT AND H. PESOTAN

Department of Mathematics, Statistics and Computer Science
University of Illinois at Chicago

AND

Department of Mathematics and Statistics
University of Guelph

Technical Report No. 84-5
Statistical Laboratory

AUGUST 1984

Research supported by Grant AFOSR80-170 and by Grant No. A8776 from NSERC.

CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b RESTRICTIVE MARKINGS
SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
DECLASSIFICATION / DOWNGRADING SCHEDULE	
PERFORMING ORGANIZATION REPORT NUMBER(S)	5 MONITORING ORGANIZATION REPORT NUMBER(S) AFOSR-TR- 84-0870
NAME OF PERFORMING ORGANIZATION University of Illinois	6b OFFICE SYMBOL (If applicable)
	7a. NAME OF MONITORING ORGANIZATION Air Force Office of Scientific Research
ADDRESS (City, State, and ZIP Code) Department of Mathematics, Statistics, and Computer Science, P.O. Box 4348, Chicago IL 60680	7b. ADDRESS (City, State, and ZIP Code) Directorate of Mathematical & Information Sciences, AFOSR, Bolling AFB DC 20332
NAME OF FUNDING / SPONSORING ORGANIZATION AFOSR	8b. OFFICE SYMBOL (If applicable) NM
	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER AFOSR-80-0170
ADDRESS (City, State, and ZIP Code) Bolling AFB DC 20332	10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO	PROJECT NO	TASK NO	WORK UNIT ACCESSION NO
61102F	2304	A5	

TITLE (Include Security Classification)
A COMPLETE CHARACTERIZATION OF TRIPLY BALANCED MATRICES WITH APPLICATIONS TO SURVEY SAMPLING

PERSONAL AUTHOR(S)
A. Hedayat and H. Pesotan

a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Technical	FROM ____ TO ____	AUG 84	9

SUPPLEMENTARY NOTATION

COSATI CODES		
FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Cross validation in survey sampling; stratified sampling; balanced half sample replication; triply balanced matrices; orthogonal arrays of strength 3; Hadamard matrices.

ABSTRACT (Continue on reverse if necessary and identify by block number)

An $R \times L$ triply balanced matrices arise in cross validation studies and in estimating the mean square errors of nonlinear statistics in many large scale survey samplings. It is shown that: (1) Any $R \times L$ triply balanced matrix and an orthogonal array $OA(R,L,2,3;\lambda)$ are one and the same object up to a possible notational change of the two symbols of the array to + and - respectively, (2) R is a multiple of 8 and $L \leq R/2$, and (3) The problem of the construction of $R \times L$ triply balanced matrices, $3 \leq L \leq R/2$, is completely resolved modulo the existence of Hadamard matrices of order $R/2$.

DISTRIBUTION / AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	UNCLASSIFIED

22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE (Include Area Code)	22c. OFFICE SYMBOL
MAJ Brian L. Woodruff	(202) 767- 5027	NM

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted SECURITY CLASSIFICATION OF THIS PAGE

A COMPLETE CHARACTERIZATION OF TRIPLY BALANCED MATRICES WITH APPLICATIONS TO SURVEY SAMPLING

BY A. HEDAYAT AND H. PESOTAN

UNIVERSITY OF ILLINOIS AT CHICAGO AND UNIVERSITY OF GUELPH

ABSTRACT

$R \times L$ triply balanced matrices arise in cross validation studies and in estimating the mean square errors of nonlinear statistics in many large scale survey samplings. It is shown that: (1) Any $R \times L$ triply balanced matrix and an orthogonal array $OA(R,L,2,3;\lambda)$ are one and the same object up to a possible notational change of the two symbols of the array to + and - respectively, (2) R is a multiple of 8 and $L \leq R/2$, and (3) The problem of the construction of $R \times L$ triply balanced matrices, $3 \leq L \leq R/2$, is completely resolved modulo the existence of Hadamard matrices of order $R/2$.

AMS subject classifications. Primary 62D05, Secondary 05B15
Key words and phrases. Cross validation in survey sampling, stratified sampling, balanced half sample replication, triply balanced matrices, orthogonal arrays of strength 3, Hadamard matrices.

A COMPLETE CHARACTERIZATION OF TRIPLY BALANCED MATRICES WITH APPLICATIONS TO SURVEY SAMPLING:

BY A. HEDAYAT AND H. PESOTAN
UNIVERSITY OF ILLINOIS AT CHICAGO AND UNIVERSITY OF GUELPH

1. Introduction. The technique of subsampling has been shown to be a very powerful method in cross-validation studies and in estimating the mean square errors of nonlinear statistics such as ratios, correlation coefficients and regression coefficients. The contributions of McCarthy (1966,1969,1976), Gurney and Jewett (1975), Lemeshow and Lèvy (1978), Krewski and Rao (1981) and Rao and Wu (1983) on this topic are noteworthy. Additional contributions in this area may be found in the references of the papers just mentioned.

The problem that we shall deal with here falls in the area of subsampling known as balanced half sample replication (BHSR) introduced and studied by McCarthy (1966,1969) for stratified samples. The set up is briefly as follows: The sampled population consists of L strata and we have a random sample of two primary sampling units (PSU) from each stratum. Thus, let y_{h1}, y_{h2} be the two observations related to the two PSU in stratum h, h = 1,2,...,L. Identify y_{h1} with + and y_{h2} with -. Prepare an R x L

matrix $\Delta = (\delta_{rh})$ with entries +1 or -1 such that in each column of Δ there are as many +1 as -1, and in addition any two columns of Δ are orthogonal. Now identify the L columns of Δ with L strata. Form R half subsamples of size L determined by the R rows of Δ. Thus the half subsample based on the i-th row of Δ is obtained as follows. If the (i,j) - th entry of Δ is +1 take y_{j1} in the i-th subsample, otherwise y_{j2} will be in the subsample. Let θ be the parameter of interest which is to be estimated based on the data. Let $\hat{\theta}_i$ be an estimator of θ based on the i-th half subsample and $\hat{\theta}$ an estimator of θ based on the entire 2L data points. Then, $v(\hat{\theta}) = c \sum_{i=1}^{R} (\hat{\theta}_i - \hat{\theta})^2$ is suggested as an estimator of the variance of $\hat{\theta}$ for a properly chosen constant c. Several authors have demonstrated that there are cases for which $v(\hat{\theta})$ behaves nicely and can be utilized in practice. Indeed, if $\hat{\theta}$ is a nonlinear statistic in the data, then perhaps one has no choice but to use $v(\hat{\theta})$ or some other similar statistic based on jackknifing or bootstraping the data.

Rao and Wu (1983) have done a serious analytical study of the above mentioned technique for a general nonlinear statistic $\hat{\theta}$ and made the following interesting discovery in the context of BHSR. They proved that if the matrix Δ has the additional feature that for any 3 of its columns

h, s and t, $\sum_{r=1}^{R} \delta_{rh}\,\delta_{rs}\,\delta_{rt} = 0$, $h \neq s \neq t$; $h,s,t = 1,2,\ldots,L$, then the statistic $v(\hat{\theta})$ enjoys additional statistical regularities. In the summer of 1984 J.N.K. Rao presented this result, obtained jointly with Jeff Wu, in the workshop on Efficient Data Collection held at the University of California at Berkeley. He proposed the existence and the construction of such matrices Δ as an open problem. In this paper we have solved this problem. Of course as we shall see this additional demand on the matrix Δ requires that more subsamples need to be taken than otherwise. This means that the practioner has to balance the need for more statistical regularity versus the cost for more computation.

2. Preliminaries. An $R \times L$ matrix $\Delta = (\delta_{rh})$ where $\delta_{rh} = +1$, or -1 will be called triply balanced if and only if

(i) $\sum_{r} \delta_{rh} = 0$, $h = 1,2,\ldots,L$,

(ii) $\sum_{r} \delta_{rh}\delta_{rs} = 0$, $h \neq s$; $h,s = 1,2,\ldots,L$,

(iii) $\sum_{r} \delta_{rh}\,\delta_{rs}\,\delta_{rt} = 0$, $h \neq s \neq t$; $h,s,t = 1,2,\ldots,L$.

Observe that (i) and (ii) imply respectively that each column of Δ is orthogonal to the vector all of whose entries are $+1$, and that any two columns of Δ are orthogonal. Condition (iii) carries no usual orthogonality

implication but is considered here since it evolved in the context of survey sampling [Rao and Wu (1983)]. The choice of the term balanced in the above definition will be justified in the next section.

In writing out matrices, + and - are used as abbreviations for +1 and -1. The existence and construction of triply balanced matrices will be taken up in Section 3. Below is an example of an 8×4 triply balanced matrix:

$$\Delta = \begin{bmatrix} + & - & + & + \\ + & + & - & + \\ - & + & - & - \\ - & - & + & - \\ + & + & + & - \\ + & - & - & - \\ - & + & + & + \\ - & - & - & + \end{bmatrix}.$$

3. Characterization of Triply Balanced Matrices. It is useful to discover for which values of R and L triply balanced matrices exist and how they may be constructed. We shall see that the three conditions of the definition place some restrictions on R and L so that such matrices cannot be easily constructed.

Since $L \geq 3$, condition (ii) implies that $R \equiv 0 (\text{mod } 4)$. We shall show that condition (iii) puts further restrictions on R. Indeed, it will turn out that such matrices exist only if $R \equiv 0 (\text{mod } 8)$ and $3 \leq L \leq R/2$. To establish this and its consequences we begin with some notation and

a lemma.

Let $f_{hs}(i,j)$ be the frequency with which the pair (i,j) occurs in any two distinct columns h and s of Δ, where $i,j \in \{+,-\}$. Similarly, let $f_{hst}(i,j,k)$ be the frequency with which the triple (i,j,k) occurs in any three distinct columns h,s,t of Δ, where $i,j,k \in \{+,-\}$.

LEMMA 3.1. If Δ is an $R \times L$ triply balanced matrix then

(1) $f_{hs}(i,j) = R/4$, for all choices of $h \neq s$;

(2) $f_{hst}(i,j,k) = R/8$, for all choices of $h \neq s \neq t$, and consequently $R \equiv 0 (\mathrm{mod}\ 8)$.

PROOF. (1). By condition (i), each column of Δ has $R/2$ plus ones and $R/2$ minus ones. Hence $f_{hs}(+,+) + f_{hs}(+,-) = f_{hs}(+,+) + f_{hs}(-,+) = R/2$. By condition (ii), $f_{hs}(+,+) - f_{hs}(+,-) - f_{hs}(-,+) + f_{hs}(-,-) = 0$. From these equations (1) follows.

(2). To simplify the notation let

$f_{hst}(+,+,+) = a$ $\quad$ $f_{hst}(-,+,+) = e$

$f_{hst}(+,+,-) = b$ $\quad$ $f_{hst}(-,+,-) = f$

$f_{hst}(+,-,+) = c$ $\quad$ $f_{hst}(-,-,+) = g$

$f_{hst}(+,-,-) = d$ $\quad$ $f_{hst}(-,-,-) = m$

By condition (i), $a+b+c+d=R/2$. By (1) and condition (ii) applied in turn to columns h and s, h and t, and s and t respectively we have

$a + b = R/4$	$a + c = R/4$	$a + e = R/4$
$c + d = R/4$	$b + d = R/4$	$b + f = R/4$
$e + f = R/4$	$e + g = R/4$	$c + g = R/4$
$g + m = R/4$	$f + m = R/4$	$d + m = R/4.$

From the above equations it follows that $b = c = e = m$ and $f = g = d$. Now by condition (iii),

$$a - b - b + R/2 - (a+b+c) - b + R/2 - (a+b+c) + R/2 - (a+b+c) - b = 0,$$

which simplifies to $a + 5b = 3R/4$. Since $a + b = R/4$ we conclude that $a = b = c = d = f = g = m$ and (2) follows. Since $f_{hst}(i,j,k)$ is an integer, therefore $R \equiv 0 (\mathrm{mod}\ 8)$, and this completes the proof.

Note that in each column of a triply balanced matrix Δ the number of plus ones is the same as the number of minus ones as required by condition (i). In every pair of columns of Δ, each type of pair (+,+), (+,-), (-,+) and (-,-) occurs equally frequently by part (1) of Lemma 3.1. In every three columns of Δ, each type of triple (+,+,+), (+,+,-), (+,-,+), (+,-,-), (-,+,+), (-,+,-), (-,-,+) and (-,-,-) occurs equally frequently by part (2) of Lemma 3.1. This observation justifies the term <u>balanced</u> used for such matrices. More importantly we can conclude

the following theorem.

THEOREM 3.1. Any R x L triply balanced matrix Δ is an orthogonal array $OA(R,L,2,3;\lambda)$ in R runs, L constraints, 2 symbols, strength 3 and index $\lambda = R/8$. Conversely, any $OA(R,L,2,3;R/8)$ is a R x L triply balanced matrix, subject to a possible notational change of the array to + and -.

Fortunately, the existence and construction of orthogor arrays $OA(R,L,2,3;\lambda)$ has been studied extensively in the literature. Hadamard matrices have been used to construct such arrays. For example, if H is a Hadamard matrix of order 4t, then it is well known that

$$\Delta = \begin{bmatrix} H \\ -H \end{bmatrix}$$

is an $OA(8t,4t,2,3;t)$ or as we established here a triply balanced $8t \times 4t$ matrix. For a summary article on Hadamard matrices see Hedayat and Wallis (1978). Indeed, we can establish a stronger relationship between triply balanced matrices and Hadamard matrices as indicated in the following:

PROPOSITION 3.1. A triply balanced R x R/2 matrix Δ exists if and only if a Hadamard matrix of order R/2 exists.

PROOF. The sufficiency is indicated above. To prove the necessity, without loss of generality we can put Δ in the form

$$\Delta = \left[\begin{array}{c|c} \underline{1} & C \\ \hline -\underline{1} & D \end{array}\right],$$

where $\underline{1}$ is a $R/2 \times 1$ column vector of plus ones. Any two columns of $\left[\begin{array}{c} C \\ \hline D \end{array}\right]$ are orthogonal and hence taken together with the first column of Δ implies that any two columns of C as well as of D are orthogonal. From this the result follows.

It is worth noting that from Margolin (1969) it follows that $D = -C$, that is, in the terminology of design of experiments $[-\underline{1}|D]$ is the fold-over of $[\underline{1}|C]$.

4. Discussion. In the context of sampling, for a given value of L we are interested in finding a $R \times L$ triply balanced matrix with R minimum. From Bose and Bush (1952) we know that in an $OA(R,L,2,3;\lambda)$ with $R \equiv 0(\text{mod } 8)$, we must have that $R \geq 2L$. Indeed, the lower bound on R is achieved via the construction based on Hadamard matrices mentioned above.

Let us summarize our findings. $R \times L$ triply balanced matrices arise in cross-validation studies and in estimating the mean square errors of nonlinear statistics in large scale survey sampling. Consequently, there is a need to

investigate when such matrices can be constructed so that practioners in survey sampling can utilize them in their work. We have shown here that:

1. Any $R \times L$ triply balanced matrix and an orthogonal array $OA(R,L,2,3; R/8)$ are one and the same object up to a possible notational change of the two symbols of the array to + and - respectively.

2. R is a multiple of 8 and $L \leq R/2$.

3. The problem of the construction of $R \times L$ triply balanced matrices, $3 \leq L \leq R/2$, is completely resolved modulo the existence of Hadamard matrices of order $R/2$.

In closing we might point out that in a similar fashion as in the argument given in Section 3, it can be shown that an orthogonal array $OA(R,L,2,t; R/2^t)$ with $t \geq 4$, is a t-ply balanced matrix and conversely. Here a t-ply balanced $R \times L$ matrix for $t \geq 4$ would have the obvious extension of the definition given here for the case $t = 3$, that is the case of triply balanced matrices. Whether or not such t-ply balanced matrices with $t \geq 4$ might be of use in survey sampling is yet to be seen.

REFERENCES

1. Bose, R.C. and Bush, K.A. (195-). Orthogonal arrays of strength two and three. Ann. Math. Statist. 23 508 - 524.

2. Gurney, M. and Jewett, R.S. (1975). Constructing orthogonal replications for variance estimation. J. Amer. Statist. Assoc. 71 819 - 821.

3. Hedayat, A. and Wallis, W.D. (1978). Hadamard matrices and their applications. Ann. of Statist. 6 1184 - 1238.

4. Krewski, D. and Rao, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. Ann. of Statist. 9 1010 - 1019.

5. Lemeshow, S. and Levy, P. (1978). Estimating the variance of ratio estimates in complex sample surveys with two primary units per stratum - a comparison of balanced replication and jackknife techniques. J. Statist. Comp. Simul. 8 191 - 205.

6. Margolin, B.H. (1969). Resolution IV fractional factorial designs. J.R. Statist. Soc. B, 31 514 - 523.

7. McCarthy, P.J. (1966). Replication: an approach to the analysis of data from complex surveys. Vital and Health Statistics. Ser. 2, No. 14, U.S. Government Printing Office, Washington, D.C.

8. McCarthy, P.J. (1969). Pseudoreplication: half-samples. Rev. Internat. Statist. Inst. 37 239 - 264.

9. McCarthy, P.J. (1976). The use of balanced half-sample replication in cross-validation studies. J. Amer. Statist. Assoc. 71 596 - 604.

10. Rao, J.N.K. and Wu, C.F.J. (1983). Inference from stratified samples: second order analysis of three methods for nonlinear statistics. Technical Report Series of the Laboratory for Research in Statistics and Probability. 7, Carleton University, Canada.

END

FILMED

11-84

DTIC